UNLIMITED

2

RSRE MEMORANDUM No. 4357

ROYAL SIGNALS & RADAR ESTABLISHMENT

TRIPHONE CLUSTERING IN THE ARM SYSTEM

Authors: M J Russell, K M Ponting, S R Browning,
S Downey & P Howell

PROCUREMENT EXECUTIVE,
MINISTRY OF DEFENCE,
R S R E MALVERN,
WORCS.

RSRE MEMORANDUM No. 4357

UNLIMITED

Royal Signals and Radar Establishment
Memorandum 4357

Triphone Clustering in the *ARM* System

M J Russell, K M Ponting, S R Browning, S Downey
and P Howell*
Speech Research Unit, SP4,
Royal Signals and Radar Establishment,
St. Andrews, Great Malvern, England

5th February 1990

Abstract

The use of triphones to cope with contextual effects in phoneme-HMM based speech recognition results in a huge increase in the number of parameters which must be estimated. One solution to this problem is to apply clustering techniques to the triphone set to produce a smaller set of "generalised triphones". An alternative is to use knowledge from phonetics of key factors which lead to context-related differences to define smaller but sufficient sets of context-sensitive HMMs. This paper reports an investigation of these methods in the context of the *ARM* continuous speech recognition system. Experiments confirm that the size of the triphone set can be substantially reduced by clustering with no degradation in recognition accuracy. These results are compared with the outcome of experiments using two knowledge-drive approaches. It is shown that, in this case, superior performance is obtained using the data-driven methods.

*Permanent address Department of Psychology, University College London

1 Introduction

The work described in this report was conducted at the UK Speech Research Unit as part of the Airborne Reconnaissance Mission (*ARM*) continuous speech recognition project. The aim of the *ARM* project is accurate recognition of continuously spoken airborne reconnaissance reports using a speech recognition system based on phoneme-level hidden Markov models (HMMs). The *ARM* project is described in [2]. The work described here is based on version 6 of the *ARM* system [2].

The more recent versions of the *ARM* system use triphone HMMs to model the context-sensitivity of the acoustic patterns corresponding to phonemes. This approach makes the simplifying assumption that context-related variations in the acoustic realisation of a particular phoneme depend only on the immediately preceding and following phonemes. This means that rather than modelling a phoneme using a single HMM, each phoneme is modelled using a set of HMMs, one for each pair of phonemes which occur as its immediate neighbours in the *ARM* baseform dictionary.

Depending on the speaker, there are approximately 1500 word-internal triphones in the *ARM* vocabulary. If, as in the present experiments, a 26 dimensional parameterisation of the acoustic front-end is used, this results in a speech recognition system with approximately 234,000 parameters. Assuming that 20 minutes of speech is used to train the system and that the acoustic front-end produces 100 frames per second, the number of training observations is 3,120,000, or approximately 13 observations per parameter. These observations are not statistically independent, nor are they uniformly distributed between triphones. In fact approximately 400 of the triphones in the *ARM* vocabulary are not represented in the training set. Consequently many of the triphone HMM parameters will be undertrained.

The solution to this training problem is to reduce the number of independent system parameters so that those which remain can be estimated more robustly from the training data. The most obvious way to achieve this is to "tie" together different system parameters so that they share the same training material. The simplest example of such an approach is the "grand" variance method [3] in which all HMM state output probability density functions share the same covariance matrix. The results of applying the grand variance method in the context of the *ARM* system are reported in [4].

An alternative method for reducing the number of system parameters is to identify classes of pairs of phonemes which have the same contextual effect on a given phoneme, and then to model the resulting equivalence class of triphones using a single HMM. The most common approach to identifying such classes is to apply clustering techniques to the full triphone set to produce clustered or "generalised" triphones [5]. An alternative to this data-driven method is to use phonetic knowledge to identify the most important factors which lead to context related differences, and then to use this information to define smaller but sufficient sets of context sensitive HMMs.

This research note reports the results of continuous speech recognition experiments using both data-driven and knowledge-driven approaches to triphone clustering. The data driven approach uses the triphone clustering algorithm described in [6] to incrementally reduce the size of the triphone set from 1500 down to less than 100 generalised triphones. In addition, two knowledge driven approaches are reported: separate modelling of syllable initial and syllable final consonants [7], and generalised triphones defined by places of articulation of the neighbouring phonemes. The corresponding HMM set sizes are 71 and 400 respectively.

This report uses the standard SAMPA computer compatible European phonetic notation system described in [10].

2 The Triphone Based *ARM* system (*ARM-6*)

The version of the *ARM* system which is used in the present experiments is *ARM-6*, a triphone-HMM based system with grand variance.

Front-end acoustic analysis in all versions of the *ARM* system is derived from the SRUbank filterbank analyser in its default configuration of 27 critical band filters spanning the range 0 to 10kHz and producing 100 frames per second. In the present experiments the feature vector $\vec{o_t} = ({o_t}^1, ..., {o_t}^{26})$ at time t is a 26 dimensional vector obtained from the SRUbank output vector $\vec{v_t}$ as follows: The mean channel amplitude $m(\vec{v_t})$ of $\vec{v_t}$ is subtracted from each component of $\vec{v_t}$ and the resulting vector is rotated using a discrete cosine transform to obtain a new feature vector $\vec{w_t}$. The vector $\vec{o_t}$ is then defined by

$$
\begin{aligned}
{o_t}^d &= {w_t}^d,\ d = 1, ..., 12 \\
{o_t}^{13} &= m(\vec{v_t}) \\
{o_t}^d &= ({w_{t+2}}^d - {w_{t-2}}^d),\ d = 14, ..., 25 \\
{o_t}^{26} &= (m(\vec{v_{t+2}}) - m(\vec{v_{t-2}}))
\end{aligned}
$$

This is the $CC12\delta$ parameterisation which was described and evaluated in [8].

Acoustic-phonetic processing in *ARM-6* uses a set of approximately 1500 HMMs (the precise number depends on the speaker) consisting of:

- Four single state "non-speech" HMMs to cope with non-speech sounds in regions of the test data between spoken sentences.
- Six word-level HMMs for the commonly occuring short words "air", "at", "in", "of", "oh" and "or". The number of states in each of these word-level HMMs

is equal to three times the number of phonemes in the baseform transcription of the corresponding word.

- Approximately 1490 three-state triphone HMMs, one for each word-internal triphone which occurs in the *ARM* vocabulary. Since the baseform pronunciations of *ARM* vocabulary words vary between speakers in the speaker-dependent *ARM* system, the precise number of triphone HMMs will be different for each speaker.

As with earlier versions of the *ARM* system, all HMM states in *ARM-6* are identified with single multivariate Gaussian state output probability density functions with diagonal (co)variance matrices. In *ARM-6* a single "grand" covariance matrix is shared by all HMM states [3].

Words in the *ARM* vocabulary are related to phonemes through a dictionary of "baseform" phonemic transcriptons. In the current, speaker-dependent, version of the *ARM* system this dictionary is modified for each speaker. These modifications are concerned with broad differences, for example between "northern english" and "southern english", rather than with fine details of the speakers pronunciation. It is assumed that spoken examples of vocabulary words conform to these baseform transcriptions.

3 HMM Training and Recognition

3.1 Training and Test Data

Speaker dependent recognition experiments were conducted using speech from three speakers (SJ, RM and MR) as training and test material. The training set consisted of 37 *ARM* reports per speaker, (224 sentences, 1985 words per speaker) chosen to give maximum coverage of phonemes which occur infrequently in the *ARM* vocabulary. Ten reports from the same speakers (540 words, 2293 phonemes per speaker) were used as test material.

3.2 Monophone HMM Training

For each speaker initial estimates of the parameters of context-insensitive monophone phoneme HMMs were obtained from the equivalent of two *ARM* reports of speech, hand labelled at the phoneme level. Similarly, initial estimates of the common word HMM parameters were obtained from single examples of these words extracted from continuous speech. The initial estimates of parameters of a single state "non-speech" HMM were obtained from a typical non-speech region of the training data.

the training data. This model was used as the initial model for all four non-speech HMMs. The models were optimised with respect to the complete training set for that speaker labelled orthographically at the sentence level. Standard sub-word HMM training procedures were used in which sentence level HMMs were constructed from phoneme-level HMMs using the dictionary of baseform transcriptions of *ARM* vocabulary words. These models were then mapped onto the sentence level acoustic data using the forward backward algorithm to obtain contributions to the model parameter estimates.

3.3 Triphone HMM Training

The parameters of the context insensitive monophone HMMs were used as the initial estimates for the parameters of the set of triphone HMMs. The triphone HMMs were then optimised with respect to the complete speaker dependent training sets labelled orthographically at the sentence level using the standard sub-word HMM training procedures. This was followed by a further three iterations of the training algorithm: the first to estimate the grand diagonal (co)variance matrix, the second to reestimate the mean vectors of the state output probability density functions given the grand (co)variance matrix, and the third to do a final reestimation of the grand (co)variance matrix. During these final three stages of training all other parameters were fixed. This "fine tuning" of the grand covariance matrix was shown to be beneficial in [4].

3.4 Recognition

Recognition was performed using a one-pass dynamic programming algorithm with beam search and partial traceback [1]. Results are presented in terms of *% words (or phonemes) correct* and *% word (or phoneme) accuracy*. These are computed as follows, using dynamic programming to align the true transcription of the test data with the output of the recogniser:

$$\% \, words \; correct = \frac{N - S - D}{N} \times 100,$$
$$\% \, word \; accuracy = \frac{N - S - D - I}{N} \times 100$$

where N is the number of words in the test set, and S, D and I are the number of words recognised as the incorrect word, deleted and inserted respectively.

Three different syntaxes were used to constrain the recognition process: a *word* syntax, which allows recognition of any sequence of words from the ARM vocabulary; a *full* syntax (perplexity 6) which was used to generate the ARM reports, and a triphone based *simple* syntax which allows any sequence of triphones to be recognised

4 Data-Driven Triphone Clustering

The data-driven approach to triphone clustering is taken from [6]. Starting with the full triphone set, the process begins by computing the distance $d(p_1, p_2)$ between all pairs of triphone HMMs p_1 and p_2 which correspond to the same phoneme. The full set of distances is then searched to find the pair of triphone HMMs (p_1, p_2) for which $d(p_1, p_2)$ is smallest. These two HMMs are then averaged to produce a single new triphone HMM and the original two triphone HMMs are discarded. This process is repeated until either the size of the triphone set has been reduced by a prespecified amount (this will be refered to as the *triphone reduction factor*), or the minimum distance is greater than some threshold.

The distance $d(p_1, p_2)$ between triphone HMMs p_1 and p_2 is define by

$$d(p_1, p_2) = \frac{avg_s \lambda_{p_1,p_2,s} + 0.5 \times max_s \lambda_{p_1,p_2,s}}{1.5}$$

where,

$$\lambda_{p_1,p_2,s} = \frac{n_1 n_2}{n_1 + n_2} \sum_{d=1}^{26} \frac{(\vec{\mu}^1_{s,d} - \vec{\mu}^2_{s,d})^2}{\sigma_d}$$

In the above expression, n_i is the number of occurrences of the triphone p_i in the training data, $\mu^i_{s,d}$ is the d^{th} component of the mean vector $\vec{\mu}^i_s$ of the s^{th} state of triphone HMM p_i, $(i = 1, 2)$, and σ_d is the d^{th} component of the grand diagonal (co)variance matrix.

For each of the speakers a full set of approximately 1500 triphone HMMs was created using the training procedures described in section 3. The clustering method described above was then applied with triphone reduction factors of 400, 500, 600, 800, 1000, 1200 and 1400. The parameters of the resulting sets of generalised triphones were then further reestimated as described in section 3.3.

5 Knowledge Driven Triphone Clustering

Two knowledge-driven approaches to triphone clustering were considered: separate modelling of syllable initial and syllable final consonants [7], and a scheme in which pairs of triphones whose contexts have the same places of articulation are assigned to the same cluster.

5.1 Separate Modelling of Syllable-Initial and -Final Consonants

The background to the work in this subsection is described in more detail in [7].

All occurrences of consonants in the *ARM* dictionary of baseform transcriptions were designated as syllable-initial or syllable-final according to a technique for location of syllable boundaries based on that described by Clements and Keyser in [9]. This method is called the *onset first principle* and has two parts, only the first of which is relevant to the current task. This part of the principle states that "Syllable-initial consonants are maximised to the extent consistent with the syllable structure conditions of the language in question". In practice what this means is "put as many consonants as are permissible before a vowel". The permissible syllable initial consonants are listed in appendix A.

Position	Phonemes
Initial	tS , dZ. b, d. D, f, g, h, j. k, l
	m, n. p, r, s, S. t, T, v, w, z.
Final	tS, dZ, b, d, D, f, g, k, l, m, n
	N, p, r, s, S, t, T, v, z, Z

Table 1: Syllable-initial and syllable-final consonants occurring in the *ARM* vocabulary for speaker MR.

Application of this technique to the *ARM* dictionaries results in the identification of 43 syllable-initial or -final consonants. These are listed for speaker MR in table 1. The corresponding lists for speakers SJ and RM are the same except that "D" does not occur in the syllable-final position for these speakers. Taken together with the vowel, common short word, and non-speech models, this results in sets of 72 HMMs for speaker MR and 71 HMMs for the other two speakers. Initial estimates of parameters for these models were obtained from monophone HMMs, trained using the method described in section 3.2. The full model set was then optimised using a further 3 iterations of the training algorithms. Because the number of HMMs in these model sets does not represent a significant increase over the size of the original monophone HMM set, grand variance is not used in this experiment.

5.2 Generalised Triphones Defined by Places of Articulation of Contexts

Each phoneme which occurs in the *ARM* vocabulary was classified according to whether its place of articulation is front (corresponding to labial or dental place of articulation, and denoted *F*), back (velar, denoted by *B*) or center (alveolar or palato-alveolar, denoted by *C*). Diphthongs are allocated two classes, corresponding to initial and final places of articulation. The classes are shown in table 2.

Phoneme	PoA	Phoneme	PoA	Phoneme	PoA
aI	CF	tS	C	dZ	C
eI	FF	A	C	{	C
oI	BF	Q	B	O	B
aU	CB	E	F	@	C
@U	CB	3	C	i	F
IU	FB	I	F	u	B
I@	FC	U	B	V	C
e@	CC	b	F	d	C
U@	BC	D	F	f	F
g	B	h	C	j	B
k	B	l	C	m	F
n	C	N	B	p	F
r	C	s	C	S	C
t	C	T	F	v	F
w	F	z	C	Z	C
?	B				

Table 2: Classification of phonemes according to place of articulation. The phonemes are represented using standard SAMPA computer readable symbols

Starting with the full triphone set, all triphones for a given phoneme which have left-contexts with a common place of articulation and right-contexts with a common place of articulation are assigned to the same cluster. Thus the triphones $(aI : n\ k)$ (denoting *aI* with left-context *n* and right context *k*) and $(aI : r\ g)$ (denoting *aI* with left-context *r* and right-context *g*) are assigned to the same cluster because *n* and *r* share the same place of articulation (C), and *k* and *g* share the same place of articulation (B). All triphones in a given cluster are then averaged, resulting in a set of approximately 400 generalised triphones. These triphones are then optimised using the procedure described in 3.3 above.

6 Experiments and Results

For reference, results of recognition experiments using monophone HMMs with the training and test data described in section 3 are presented in table 3. The state output pdfs of the HMMs used in these experiments have state-specific (i.e. not grand) covariance matrices.

	Phoneme Syntax (Perplexity=53)		Word Syntax (Perplexity=497)		Full Syntax (Perplexity=6)	
Speaker	Phonemes Correct	Phoneme Accuracy	Words Correct	Word Accuracy	Words Correct	Word Accuracy
MR	65.6%	49.9%	78.3%	53.3%	98.1%	97.8%
RM	63.4%	48.1%	73.5%	42.8%	98.0%	96.5%
SJ	66.2%	53.3%	79.8%	52.2%	99.1%	98.7%
Average	65.1%	50.4%	77.2%	49.4%	98.4%	97.7%

Table 3: Results of experiments using context-insensitive monophone HMMs (540 word test set per speaker).

6.1 Experiments with Data Driven Triphone Clustering

The results of recognition experiments using sets of generalised triphones defined using the data-driven triphone clustering method of section 4 are summarised in figure 1. The figure shows % word accuracy with no syntax as a function of triphone reduction factor for the three speakers. The results show no significant drop in performance until the size of the triphone set is reduced to less than 300. This is consistent with the results of triphone clustering experiments reported in [5].

Note that performance with a triphone reduction factor of 400 is guaranteed to be exactly the same as performance with no clustering. This is because over 400 of the triphones in the *ARM* system are not represented in the training set. Under the current training scheme, these HMMs retain the relevant monophone HMM statistics. Therefore they are clustered into a single generalised triphone, and this generalised triphone has identical parameters to the original monophone.

The complete results for these experiments are presented in appendix B.

6.2 Experiments with Knowledge Driven Triphone Clustering

6.2.1 Separate Modelling of Syllable-Initial and -Final Consonants

The results of recognition experiments using separate HMMs for syllable-initial and syllable-final consonants are shown in table 4. The final row of the table shows average results for the corresponding number of generalised triphones derived using the data-driven method.

The results are significantly better than those for monophone HMMs shown in table 3: The average word accuracy with no syntax is raised from 49.4% for

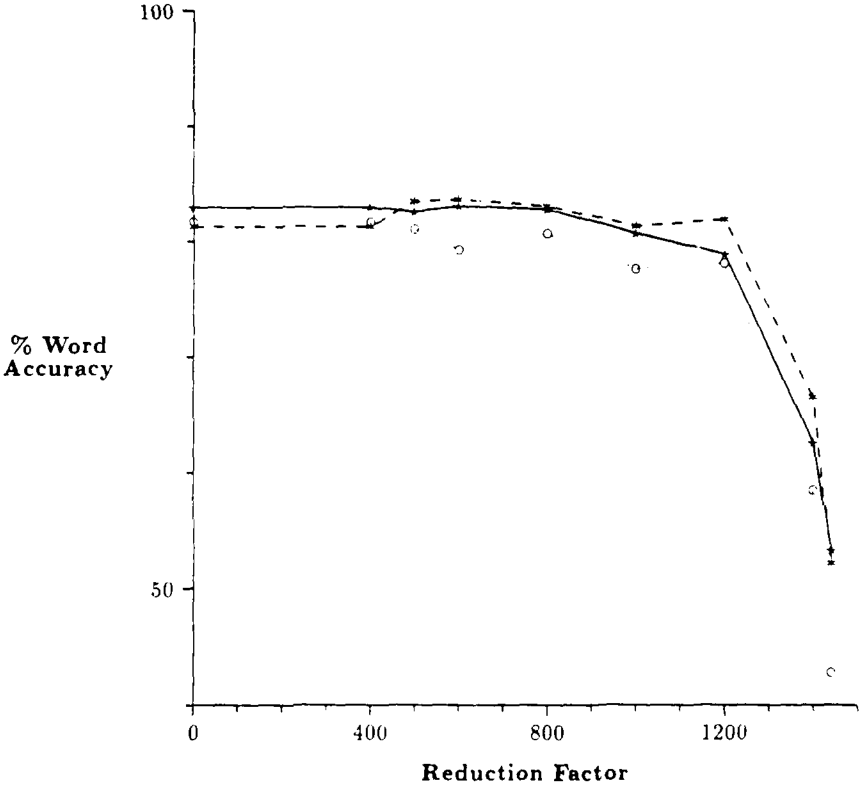

Figure 1: Word accuracy as a function of triphone reduction factor for speakers *MR* *(solid line), RM (dotted line) and SJ (dashed line) with no syntax.*

	Phoneme Syntax (Perplexity=72)		Word Syntax (Perplexity=497)		Full Syntax (Perplexity=6)	
Speaker	Phonemes Correct	Phoneme Accuracy	Words Correct	Word Accuracy	Words Correct	Word Accuracy
MR	66.4%	51.3%	81.5%	59.6%	98.1%	97.8%
RM	60.3%	43.3%	79.6%	56.1%	99.3%	98.5%
SJ	67.9%	54.1%	85.0%	62.0%	98.5%	97.6%
Average	64.7%	49.6%	82.0%	59.2%	98.6%	98.0%
Average for 71 G Triphones	69.5%	57.0%	85.7%	67.2%	99.3%	99.0%

Table 4: Results of experiments using separate modelling of syllable-initial and -final consonants (540 word test set per speaker).

context-insensitive HMMs to 59.2% when syllable-initial and -final consonants are modelled separately. However the results compare badly with the figures for the same number of generalised triphones derived using data-driven clustering. For example, the average word accuracy with no syntax show in table 4 for separate modelling of syllable-initial and -final consonants is 59.2% compared with 67.2% for the same number of generalised triphones.

6.2.2 Place of Articulation Triphones

Table 5 shows the results of recognition experiments using generalised triphones defined by places of articulation of context, as described in section 5.2. For comparison, the final row of the table shows the average results over 300 and 500 generalised triphones derived using the data-driven method.

As was the case with separate modelling of syllable-initial and -final consonants, the results for "place of articulation" triphones are significantly better than those for context-insensitive monophone HMMs, but compare badly with results which were obtained using data-driven triphone clustering methods to derive similar numbers of generalised triphones. For example, table 5 shows that the average word accuracy with no syntax for the 400 "places of articulation" triphones is 74.5%, compared with 79.6% and 79.9% respectively for sets of 300 and 500 generalised triphones derived using data-driven clustering.

	Phoneme Syntax		Word Syntax (Perplexity=497)		Full Syntax (Perplexity=6)	
Speaker	Phonemes Correct	Phoneme Accuracy	Words Correct	Word Accuracy	Words Correct	Word Accuracy
MR	62.6%	44.9%	88.5%	75.0%	98.9%	98.3%
RM	56.6%	38.8%	87.8%	69.3%	98.0%	96.3%
SJ	59.9%	44.8%	92.6%	79.1%	98.9%	98.5%
Average	59.7%	42.8%	89.6%	74.5%	98.6%	97.7%
Average over 300 and 500 G Triphones	73.9%	52.3%	92.6%	79.7%	99.4%	98.9%

Table 5: Results of place of articulation triphone clustering experiments (540 word test set per speaker).

6.3 Summary of Results

The results of all of the experiments reported in section 6 are summarised in figure 2. The figure shows % word accuracy with no syntax averaged over the 3 speakers.

7 Conclusions

The results presented in this report demonstrate that the number of HMMs in a triphone HMM based speech recognition system can be substantially reduced with no significant reduction in speech recognition accuracy.

The standard data-driven approach to triphone clustering has been compared experimentally with two particular knowledge driven approaches to modelling contextual effects: separate modelling of syllable-initial and -final consonants and "place of articulation" triphones. In this case, the results do not support the hypothesis that knowledge driven approaches to modelling contextual effect are advantageous over the data-driven generalised triphone method. In fact, the superior performance of the 71 generalised triphones over the same sized model set derived by separate modelling of syllable-initial and -final consonants is statistically significant.

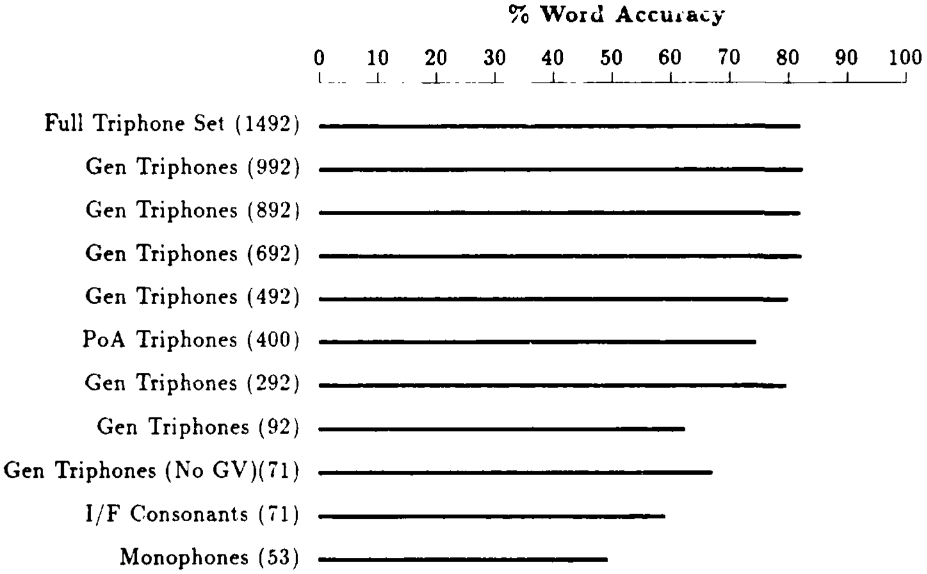

Figure 2: Bar chart summarising the results presented in the previous section. The figure shows average % word accuracy with no syntax for all HMM sets considered in the experiments. The figures in brackets are the sizes of the HMM sets.

References

[1] J S Bridle, M D Brown and R M Chamberlain, "A one-pass algorithm for connected word recognition", IEEE-ICASSP, 899-902, 1982.

[2] M J Russell, K M Ponting, S M Peeling, S R Browning, J S Bridle and R K Moore, "The ARM Continuous Speech Recognition System", Proc. ICASSP'90, Albuquerque, New Mexico, April 1990.

[3] D B Paul, "A speaker-stress resistant isolated word recognizer", ICASSP'87, Dallas, TX, 1987.

[4] M J Russell and K M Ponting, "Experiments with Grand Variance in the ARM Continuous Speech Recognition System", RSRE Memorandum Number 4359, 1990.

[5] K-F Lee, "Large Vocabulary Speaker-Independent Continuous Speech Recognition: the SPHINX System", PhD Thesis, Carnegie Mellon University, 1988.

[6] D B Paul, "Speaker-stress resistant continuous speech recognition", Proc ICASSP'88, New York, 1988.

[7] P Howell, "Phone models for an automatic speech recognition system based on hidden Markov models", SP4 Research Note Number 87, RSRE, September 1989.

[8] M J Russell, D Lowe, M D Bedworth and K M Ponting. "Improved Front-End Analysis in the ARM System: Linear Transformations of SRUbank", RSRE Memorandum Number 4358, 1990.

[9] G K Clements and S J Keyser, "CV Phonology: A Generative Theory of the Syllable", MIT Press, Cambridge Massachusetts, 1983.

[10] J Wells et al., "Specification of SAM Phonetic Alphabet (SAMPA)", included in: P Winski, W J Barry and A Fourcin (Eds), "Support Available from SAM Project for other ESPRIT Speech and Language Work", The SAM Project, Department of Phonetics, University College London.

A Permissible syllable-initial consonants and consonant clusters

The following infoprmation is based on [9] and [7].

Single Consonants: All consonants except N are permitted initially.

Pairs of Consonants:

	w	l	r	p	t	k	m	n	f	T	j
p	-	+	+	-	-	-	-	-	-	-	+
b	-	+	+	-	-	-	-	-	-	-	+
f	-	+	+	-	-	-	-	-	-	-	+
v	-	-	-	-	-	-	-	-	-	-	+
t	+	-	+	-	-	-	-	-	-	-	+
d	+	-	+	-	-	-	-	-	-	-	+
T	+	-	+	-	-	-	-	-	-	-	+
h	-	-	-	-	-	-	-	-	-	-	+
k	+	+	+	-	-	-	-	-	-	-	+
g	+	+	+	-	-	-	-	-	-	-	+
l	-	-	-	-	-	-	-	-	-	-	+
s	+	+	-	+	+	+	+	+	-	-	+
S	+	+	+	-	-	-	-	-	-	-	-
m	-	-	-	-	-	-	-	-	-	-	+
n	-	-	-	-	-	-	-	-	-	-	+

Triples of consonants:

	w	l	r	j
sp	-	+	+	+
st	-	-	+	+
sk	+	+	+	+

For pairs and triples, rows specify the first member of the clusters and the columns specify the second member. A "+" indicates that the row/ column pair is a permitted syllable-initial cluster while a "-" indicates that it is not.

B Results of triphone clustering experiments

This appendix presents tables which show the complete results of the data-driven triphone clustering recognition experiments for the three speakers MR, RM and SJ. The final row of each table is the figure for context-insensitive monophone HMMs. The penultimate row of each table, marked with the symbol "†", corresponds to a generalised triphone set where the number of models is the same as in the knowledge-driven "separate modelling of syllable-initial and -final consonants" experiment. For consistency with the latter, the generalised triphone sets which gave rise to the results in the columns marked †did not use grand variance. This explains why the figures in these rows are superior to those in the immediately neighbouring rows.

	Phoneme Syntax		Word Syntax (Perplexity=497)		Full Syntax (Perplexity=6)	
Reduction Factor	Phonemes Correct	Phoneme Accuracy	Words Correct	Word Accuracy	Words Correct	Word Accuracy
0-400	69.8%	39.4%	94.8%	81.3%	99.4%	99.1%
500	72.6%	46.1%	95.2%	83.5%	99.3%	98.7%
600	73.4%	47.4%	95.2%	83.7%	99.3%	98.7%
800	73.8%	50.0%	95.0%	83.1%	99.6%	99.4%
1000	73.4%	51.6%	95.0%	81.3%	99.3%	98.7%
1200	73.8%	54.4%	93.5%	81.9%	99.3%	98.7%
1400	66.3%	55.3%	88.0%	66.5%	99.4%	99.1%
1419†	69.4%	58.1%	88.1%	69.8%	99.4%	99.4%
1434	66.2%	53.3%	79.8%	52.2%	99.1%	98.7%

Table 6: Results of data-driven triphone clustering experiments (speaker SJ, 540 word test set).

	Phoneme Syntax		Word Syntax (Perplexity=497)		Full Syntax (Perplexity=6)	
Reduction Factor	Phonemes Correct	Phoneme Accuracy	Words Correct	Word Accuracy	Words Correct	Word Accuracy
0-400	71.0%	40.4%	94.1%	81.7%	99.4%	98.7%
500	68.9%	38.6%	93.9%	81.1%	99.4%	98.7%
600	72.4%	44.6%	93.0%	79.3%	99.3%	98.3%
800	72.9%	47.1%	93.9%	80.7%	99.4%	98.7%
1000	72.6%	48.2%	92.4%	77.6%	99.6%	99.1%
1200	72.1%	50.0%	91.5%	78.1%	99.6%	99.3%
1400	64.3%	49.7%	83.3%	58.5%	99.3%	98.5%
$1424^{\dagger}$	66.6%	52.7%	85.6%	66.3%	99.3%	98.9%
1439	63.4%	48.1%	73.5%	42.8%	98.0%	96.5%

Table 7: Results of data-driven triphone clustering experiments (speaker RM, 540 word test set).

	Phoneme Syntax		Word Syntax (Perplexity=497)		Full Syntax (Perplexity=6)	
Reduction Factor	Phonemes Correct	Phoneme Accuracy	Words Correct	Word Accuracy	Words Correct	Word Accuracy
0-400	73.4%	45.1%	94.1%	83.0%	99.8%	99.8%
500	75.6%	50.9%	93.5%	82.6%	99.8%	99.8%
600	76.0%	51.4%	93.3%	83.1%	99.8%	99.8%
800	76.7%	53.2%	92.8%	82.8%	99.6%	99.4%
1000	76.1%	54.2%	91.9%	80.7%	99.4%	99.1%
1200	75.5%	55.1%	91.5%	78.9%	99.3%	98.7%
1400	-	-	85.0%	62.6%	-	-
$1421^{\dagger}$	72.5%	60.3%	83.3%	65.6%	99.3%	98.7%
1436	65.6%	49.9%	78.3%	53.3%	98.1%	97.8%

Table 8: Results of data-driven triphone clustering experiments (speaker MR, 540 word test set).

DOCUMENT CONTROL SHEET

Overall security classification of sheet ....Unclassified..

(As far as possible this sheet should contain only unclassified information. If it is necessary to enter classified information, the box concerned must be marked to indicate the classification, eg (R), (C) or (S))

1. DRIC Reference (if known)	2. Originator's Reference MEMO 4357	3. Agency Reference	4. Report Security Classification Unclassified

5. Originator's Code (if known) 7784000	6. Originator (Corporate Author) Name and Location Royal Signals & Radar Establishment St Andrews Road, Great Malvern Worcestershire WR14 3PS
5a. Sponsoring Agency's Code (if known)	6a. Sponsoring Agency (Contract Authority) Name and Location

7. Title

TRIPHONE CLUSTERING IN THE ARM SYSTEM

7a. Title in Foreign Language (in the case of Translations)

7b. Presented at (for Conference Papers): Title, Place and Date of Conference

8. Author 1: Surname, Initials RUSSELL M J	9a. Author 2 and others	9b Authors 3, 4	10. Date 1990.2	pp. ref 16

11. Contract Number	12. Period	13. Project	14. Other Reference

15. Distribution Statement

Unlimited

Descriptors (or Keywords)

Continue on separate piece of paper

Abstract

The use of triphones to cope with contextual effects in phoneme-HMM based speech recognition results in a huge increase in the number of parameters which must be estimated. One solution to this problem is to apply clustering techniques to the triphone set to produce a smaller set of "generalised triphones". An alternative is to use knowledge from phonetics of key factors which lead to context-related differences to define smaller but sufficient sets of context-sensitive HMMs. This paper reports an investigation of these methods in the context of the 'ARM' continuous speech recognition system. Experiments confirm that the size of the triphone set can be substantially reduced by clustering with no degradation in recognition accuracy. These results are compared with the outcome of experiments using two knowledge-drive approaches. It is shown that, in this case, superior performance is obtained using the data-driven methods.